Computer Science
TEXT REPRESENTATION ON LITMUS PAPER

Philip R. Johnson and Beomjin Kim*

Department of Computer Science
Indiana University – Purdue University Fort Wayne
2101 E. Coliseum Blvd
Fort Wayne, Indiana 46805

The advent of the information age leaves us stranded in an ocean of data. Finding relevant information from this vast information workspace is a difficult and time-consuming process. Even a well-focused query can easily produce an overwhelming number of hits. Providing efficient and interactive tools to assist users in their navigation of this information workspace is indispensable. Researchers have proposed various text-retrieval systems to support users in the search process. Current text-retrieval systems, mainly based on text summarization or document ranking, should be improved to meet the users' expectations more adherently.

In this research, we have developed a text representation system (TRS) that transforms text-based information into a visualized format and provide interactive tools for easy access and interpretation of the underlying texts. A text is divided into multiple paragraphs that are projected as a sequence of colored strips. Given queries are mapped into a primary color, and the frequency of each query in a paragraph determines the intensity of the corresponding query's color. The paragraph length determines the width of the strip onto which the mixture of colors is painted. This general view of a text allows users to compare multiple texts intuitively and swiftly. The TRS provides zoom-in and filtering functionality for further navigation in the analysis of a specific section of the text on demand. The system also suggests other significant terms for a paragraph and the entire text. These terms can then be applied to refine their query for concentrating or expanding the search scope when an unexpected search result appears. The TRS has been implemented on a J2EE environment, and interactive navigation tools are constructed with Java swing components.

A pioneer study has been conducted on texts in the *New York Times* archive in which the subjects have shown a positive reaction with a high degree of satisfaction using our TRS. In summary, the TRS simultaneously shows paragraph lengths, term occurrences, query distribution, and relationships between queries. This compact, but informative visual representation supports decision processes, which will alleviate the referencing of text. This TRS can be immediately applicable to project a large volume of text information, including Internet search results and digital library systems. Our on-going research is under expansion to be function in conjunction with an Internet search engine, which will eventually save users' time and decrease network traffic by reducing the number of references to unrelated texts.